

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
5 December 2002 (05.12.2002)

PCT

(10) International Publication Number
WO 02/098059 A1

(51) International Patent Classification⁷: H04L 12/24, 12/56

(21) International Application Number: PCT/FI02/00224

(22) International Filing Date: 19 March 2002 (19.03.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
20011114 28 May 2001 (28.05.2001) FI

(71) Applicant (for all designated States except US): NOKIA CORPORATION [FI/FI]; Keilalahdentie 4, FIN-02150 Espoo (FI).

(72) Inventor; and

(75) Inventor/Applicant (for US only): SAKSIO, Mauri [FI/FI]; Puosunrinne 4 B 30, FIN-02320 Espoo (FI).

(74) Agent: PAPULA OY; P. O. Box 981, (Fredrikinkatu 61 A), FIN-00101 Helsinki (FI).

(81) Designated States (*national*): AE, AG, AL, AM, AT (utility model), AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ (utility model), CZ, DE (utility model), DE, DK (utility model), DK, DM, DZ, EC, EE (utility model), EE, ES, FI (utility model), FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK (utility model), SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

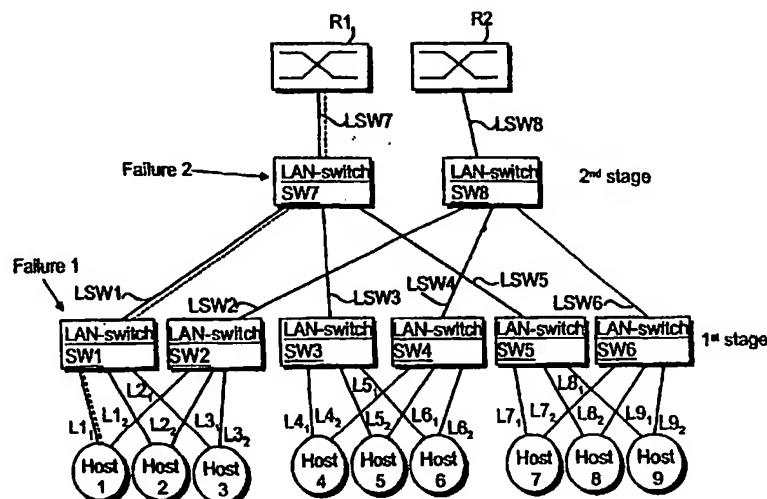
(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHOD AND SYSTEM FOR IMPLEMENTING A FAST RECOVERY PROCESS IN A LOCAL AREA NETWORK



(57) Abstract: The present invention concerns a method and a system for accelerating fault recovery in a redundant, tree structured local area network. The invention is used to define some of the LAN ports, which are, for example, used to connect the switch (SW) into the IP router, as critical ones. Likewise, some other LAN ports, used to connect the IP hosts to said switch (SW), are defined as dependent of the critical links. If a critical LAN port or corresponding link is found to be non-functional, e.g. no carrier sensed, all LAN ports or corresponding links depending on it are declared as non-functional. The declaration is done at link level in a way which allows the device(s) or ports connected to the other end of the link to notice that the link is not in use anymore to carry traffic.

METHOD AND SYSTEM FOR IMPLEMENTING A FAST RECOVERY PROCESS IN A LOCAL AREA NETWORK

FIELD OF THE INVENTION

The present invention relates to local area networks (LAN). In particular, the present invention relates to a novel and improved method and system for implementing a fast recovery process in a local area network.

BACKGROUND OF THE INVENTION

The local area network (LAN) is a group of computers and associated devices that share a common communications line and typically share the resources of a single processor or server within a small geographic area (for example, within an office building, within certain parts of IP backbone networks or within a network element, such as a telephone exchange or network control element). In this context, the local area network can also mean an architecture that uses a so-called "loosely coupled multiprocessor" architecture and in which the messages between the processors are sent via Ethernet. This kind of architecture can be implemented, for example, in the IP trunks, MSC Servers (Mobile Switching Center, MSC) or network elements such as Connection Processing Server (CPS) or Home Subscriber Server, which are used in 'all IP' architectures of the third generation mobile networks.

Usually, the server has applications and data storage that are shared in common by multiple computer users or central processor units (CPU's). The local area network may serve as few as two or three users or clients (for example, in a home network) or as many as thousands of users. The information is transmitted between two clients or hosts using the paths from the first client to the second client. These paths are formed using the links between the two network ele-

ments. Typically the paths are formed beforehand. In redundant networks the first link of the host or client to the first network node is duplicated, thus allowing a recovery to the other path in a fault situation of the first path.

A router is a device or, in some cases, software in a computer that determines the next network point to which a packet should be forwarded toward its destination. The router is connected to at least two networks and it decides which way to send each information packet based on its current understanding of the state of the networks it is connected to. The router is often included as a part of a network switch.

The switch is a network device that selects a path or circuit for sending a unit of data to its next destination. The switch may also include the function of the router, a device or program that can determine the route and specifically what adjacent network point the data should be sent to. In general, a switch is a simpler and faster mechanism than a router, which requires knowledge about the network and how to determine the route.

Relative to the layered Open Systems Interconnection (OSI) communication model, a switch is usually associated with Layer 2, the Data Link Layer. However, some newer switches also perform the routing functions of layer 3, the Network Layer. Layer 3 switches are also sometimes called IP switches.

On computer and telecommunication devices, a port is generally a specific place for being physically connected to some other device, usually with a socket and plug of some kind. A link is a physical and, in some usage, a logical connection between two points. Both ends of the link are usually connected to the port.

In this context, the term "host" means any computer that has a complete two-way access to other computers in the network. A host has a specific "local or host number" that, together with the network number, forms its unique address. A "host" is a node in a network.

To maintain the operation of the network one has to take care of the fact that all the substantial elements are operational. The management of faults as a part of the network management improves the reliability of the network, thereby providing the maintainer of the network and the network itself with the tools for promptly detecting the faults and correcting them. The responsibility of the management of faults is to arrange things so that problems and interruptions would be visible to the users as little as possible.

The devices in the network may send a notification of a critical situation every time there is a fault situation occurring (*logging*). Examples of critical situations are, e.g. the rebooting of a device or a response that was never received from a device. In most of the cases, the management of faults based on merely this kind of information does not give a sufficient picture of the state of the network. For example, when some device is damaged, it is not always able to send a notification thereof.

The devices of the network may be regularly asked about their status (*polling*). Enquiries such as this enable one to detect the faults quite promptly. However, they take the capacity of the network from the actual payload. One has to balance between the detection accuracy and network capacity to be used, i.e. the greater the detection accuracy one wishes to have, the bigger part of the transfer capacity of the network is used. Other matters that have an influence on the selection of the polling interval are the number

of the devices to be monitored and the capacity of the links to be used.

When the failure is detected, one has to accurately locate the fault and isolate the rest of the network from the disturbance caused by the fault. The network has to be configured or changed in such a way that the effects of the elimination of a component on the operation of the network are minimised. Finally, the network is reset by correcting or changing the faulty components.

However, there are situations and network solutions in which the above-mentioned methods for the management of faults and most of all for the fault detection are not applicable because the fault has to be detected without delay. For example, in the internal network structure of a network element or IP network, the failure of a link combining two plug-in units may cause problems in an ongoing call or real time data connection, in which case the fault has to be detected very fast, in order that the call or connection would not be interrupted and that the users would not detect the fault.

The standard method in a redundant local area network is to use the Spanning Tree Protocol (STP) or some vendor-specific, proprietary solution. The spanning tree protocol and algorithm were developed by a committee of the IEEE (Institute of Electrical and Electronics Engineers). Currently, the IEEE is attempting to institute enhancements to the spanning tree algorithm that will reduce network recovery time. The goal is to go from 30 to 60 seconds after a failure or change in link status to less than 10 seconds. However, due to the long recovery time needed, the STP is not suitable for environments requiring fast (a maximum of few seconds) recovery.

An alternative solution is that each IP host monitors that it has a functioning link to some criti-

cal part of the LAN (typically this is the router connecting the host to the external IP network). A simple method to implement the monitoring is to use the ICMP ECHO (ping) messages, which are sent to the router and to which it is supposed to respond. ICMP is a message control and error-reporting protocol between a host server and a gateway to the Internet. ICMP uses Internet Protocol datagram, but the messages are processed by the IP software and are not directly apparent to the application user.

The major problem of the standard STP is its possibly slow recovery (recovery may take several tens of seconds during which time part or all of the LAN won't carry traffic). Vendor-specific solutions are much faster, but they require that all critical equipment (mainly LAN switches) be purchased from a single provider.

There are also some problems with the ICMP ECHO method: It can be used only with links which have a corresponding IP address. That is, this method cannot be used if we have a redundant LAN port which does not have an IP address bound to it (for example, the port is just idling and can be used in case of a primary LAN port failure). The ICMP ECHO messages create some extra load for the LAN and especially for the router (or some other device which the host wants to ping). Thus, it is not possible to monitor the functionality of the link constantly but only intermittently, for example, once in five seconds. Some ECHO messages can also be lost due to congestion and thus the recovery can be started only after a few unanswered messages. As a result, even though the recovery itself can be very fast, the detection of a fault is still rather slow, taking from several seconds to approximately 20 seconds.

SUMMARY OF THE INVENTION

The present invention concerns a method and a system for accelerating fault recovery in a redundant, tree structured local area network. In this context, the tree structure means that there are no closed loops in the network. The tree is a directed non-cyclic network. The invention is used to define some of the LAN ports, which are, for example, used to connect the switch into the IP router, as critical ones. Likewise, some other LAN ports, used to connect the IP hosts to said switch, are defined as dependent of the critical links. If a critical LAN port or corresponding link is found to be non-functional, e.g. no carrier sensed, all LAN ports or corresponding links depending on it are declared as non-functional. The declaration is done at link level in a way which allows the device(s) or ports connected to the other end of the link to notice that the link is not in use anymore to carry traffic. Thus, the net effect is that the knowledge of the fault at the upper level of the tree is propagated very fast down to the hosts, thus enabling fast recovery.

The present invention may enable a considerably fast detection time of a failure taking about a second, perhaps even less. Because of this, the recovery time can be reduced significantly. Also the fault detection, according to the present invention, does not load the LAN, even though the load reduction is not likely to be significant. Also, the usability of the present invention does not require that there is an IP address bound to all ports (links) to be monitored.

The present invention also overcomes the problems of the ICMP ECHO mechanism in the sense that the ICMP ECHO mechanism is an end-to-end verification of the path, whereas the present invention can guarantee that the physical path from the host to the external IP network or vice versa is in use.

Also the present invention can be implemented in a way which is compatible with the current LAN switches. The reason for this is the fact that the inventive mechanism does not require any protocol between the LAN switches.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are included to provide a further understanding of the invention and constitute a part of this specification, illustrate embodiments of the invention and together with the description help to explain the principles of the invention. In the drawings:

Fig 1 is a block diagram illustrating a network structure according to one embodiment of the present invention, and

Figs 2a-2b describe a structure of the network element according to one embodiment of the present invention in more detail.

DETAILED DESCRIPTION OF THE INVENTION

Reference will now be made in detail to the embodiments of the present invention, examples of which are illustrated in the accompanying drawings.

In Figure 1 there is described a redundant LAN, which has the topology of a tree. The term "redundant" means that the host connection has been duplicated in order to allow a switch over from the active link $L1_1$ to the standby link $L1_2$ in a link or a path failure situation. Also in figure 1 there is one active connection (traffic flow) described with the dash line. This connection is established between the Host 1 and the router R1. It should be noted that the LAN topology in this example is such that there are at least two stages of LAN switches.

If one of the 1st stage LAN switches SW1, ..., SW6 has failed, failure 1, or has been powered down for maintenance etc. there is not a big problem because the hosts Host 1, ..., Host 9 are connected directly to the 1st LAN switches and they can detect themselves when a link/LAN port goes from link-up state to link-down state. The recovery can be initiated immediately when the LAN driver software in the host notifies of the link-down situation. If one of the 2nd stage LAN switches SW7, SW8 has failed, failure 2, the situation is the same if the link to the corresponding router or a link between a 1st stage and 2nd stage LAN switch has failed, the problem is that because the hosts are not directly connected to the 2nd stage LAN switch, they do not directly detect the failure. This is because the link from the host to the 1st stage LAN switch stays in the link-up state. The recovery only starts when the hosts find out that the router is not responding to the ICMP ECHO messages. But as mentioned above, this is not the best and fastest way to start the recovery process.

In the following there is described the idea of the present invention. In the failed 2nd stage LAN switch SW7 it has been defined that the link LSW7 to router R1, called up-link, is a critical link and the so-called down-links LSW1, LSW3, LSW5 to the 1st stage LAN switches SW1, SW3, SW5, are dependent of the critical up-link LSW7. Thus, if the up-link LSW7 fails, all down-links LSW1, LSW3, LSW5 are set in the link-down state. Likewise, in the 1st stage LAN switches SW1, SW3, SW5, the links LSW1, LSW3, LSW5 to the 2nd stage LAN switch SW7 are defined to be as critical and links L1₁, L2₁, L3₁, L4₁, L5₁, L6₁, L7₁, L8₁, L9₁ to the hosts 1, 2, ..., 9, down-links, are defined to be dependent of the up-links LSW1, LSW3, LSW5. The net result is that if the 2nd stage LAN switch or its link to the router fails, failure 2, then the link-

down state is propagated down to hosts Host 1, ..., Host 9. The same will happen if the link between a 1st stage and 2nd stage LAN switch fails. Thus, the hosts become very quickly aware of a failure in the LAN and can start recovery immediately.

One example of said recovery is that the host transfers to a predetermined default mode. This is the case if also the redundant up-link, e.g. link L1₂ for Host 1, is in a link down state. For instance the Home Subscriber Server, an example of possible Host 1, is solving a profile of a certain user and it needs to be connected to the other network element (not shown) behind the routers R1, R2. If both links L1₁ and L1₂ are in link down state, the recovery in this example is that Host 1 uses a predetermined default profile for said user. The only important matter is that the host is notified as soon as possible of the link down situation of said active and redundant links.

It must be noted that the necessary changes will be implemented in the LAN switches, even though co-operation with the host software is needed. The host moves all LAN traffic into the redundant LAN port if the currently used LAN port is changed into a link-down state. It must also be noted that there can be more than one critical link per LAN switch and that a link can depend on zero, one or more critical links. If a link depends on more than one critical links, the link will be put into a link-down state if any of the critical links is in a link-down state.

In the case of a failed link, the LAN switch or router is repaired and put into operation, and all ports connected to it are put into a link-up state unless otherwise specified by some management operation. As a result, all links dependent of it are also put into a link-up state unless overridden by management operations. This process is very much the same as in a

failure situation where the hosts are notified of the failure situation.

The above described inventive mechanism can also be used to notify the hosts or the LAN switches, if there is something wrong with the transmission direction of the connection. The idea is that normally a device cannot know whether or not it is transmitting properly or whether or not the receiving device is receiving properly. However, it is possible to think of a link to be dependent of itself and change the state of the link into a link-down state if it is noticed that the device on the other end of the link is not receiving or sending properly, i.e. there are excessive CRC (cyclic redundancy check) errors, runt frames etc.

In figure 2a there is described a coarse example of the LAN Switch structure according to one embodiment of the present invention. In figure 2b there is described a coarse example of the host or CPU unit structure according to one embodiment of the present invention.

In both examples there is an Ethernet controller or Ethernet physical layer transceiver EC connected to the network element itself. The Ethernet controller EC is further divided at least in two components or modules which, of course, can be in the same circuit. These modules are the Media Access Controller MAC and the physical layer device PHY. The media access layer communicates directly with the network adapter card and is responsible for delivering error-free data between two computers. The physical layer device PHY performs the same general function as a transceiver in the typical Ethernet system.

For a typical network connection the data terminal equipment, LAN switch, host or CPU device (computer) contain an Ethernet interface EC which generates and sends Ethernet frames that carry data be-

tween computers attached to the network. The interface or repeater port might also be designed to include the PHY electronics internally. In the present invention the Ethernet controller EC is designed to monitor the status of the active link. After the Ethernet controller has noticed a link-down situation, it "sends" information about the situation downwards by setting the downward links into a link-down state. When the Ethernet controller in the host notices the link-down situation of the active link it notifies the host software, and the recovery can be started.

In figure 2a there is described the implementation of N ports into one LAN-Switch. The Ethernet Controller EC comprises n pairs of the media access controller MAC - physical layer device PHY. Physical layer devices are connected to the control logic, which typically can be implemented by a microprocessor in order to monitor and control the state of the PHY devices.

The essential feature of the PHY devices is that they contain or provide an information signal and/or register that informs of the state of the link or port. It is also useful if the information can be monitored using software. Also the PHY device can provide said information by producing an interruption to the microprocessor that can interpret this interruption as a change of the state of the PHY device. Another essential feature of the PHY device is that it can be reset into the state in which it does not give idle information to the other PHY device. In figure 2a, the control of the above-mentioned two essential features is described using two different signal types. "Link Down" indication signals are sent from the PHY devices in order to inform the Control logic of the present situation of the link. Thus the PHY devices can be set into the state which can be recognised as a failure situation in the down link of said

devices. "PHY Reset" signals are used to set the PHY devices into the down state so that the other PHY device in a down link direction can recognise the failure in the up link direction, i.e. these signals disable the PHY devices.

It is obvious to a person skilled in the art that with the advancement of technology, the basic idea of the invention may be implemented in various ways. The invention and its embodiments are thus not limited to the examples described above, instead they may vary within the scope of the claims.

CLAIMS

1. A method for fast recovery of a host connection in a redundant tree structured local area network, characterised in that the method comprises the steps of:

monitoring the state of a critical up-link,
setting a dependent down-link in a link-down state, if said critical up-link is detected to be in a link-down state.

monitoring the state of a active up-link in the host device, and

starting a recovery process in a host device if said active link is in the link-down state,

2. The method according to claim 1, characterised in that specifying the up-link of a network element being a critical up-link, if the failure of said link affects the data flow of a down-link of said network element.

3. The method according to claim 1, characterised in that specifying the link of a network element being a dependent down-link, if there is a critical up-link between said down-link and the next network element.

4. The method according to claim 1, characterised in that the recovery process comprises the steps of:

notifying the host software of the link failure in the active up-link, and

changing the active data path to the redundant up-link.

5. The method according to claim 1, characterised in that the recovery process comprises the steps of:

notifying the host software of the link failure in the active up-link,

checking the status of the redundant up-link, and if said up-link is in link down state,

transferring said host to the predetermined default mode operation.

6. The method according to claims 4 or 5, characterised in that said redundant up-link is a doubling up-link for said active up-link.

7. The method according to claim 1, characterised in that monitoring the state of a critical up-link is accomplished by monitoring the quality of the data flow on the link.

8. A system for fast recovering of a host connection in a redundant tree structured local area network, characterised in that the system comprises

a monitoring device (EC) for monitoring the state of a critical up-link, for setting a dependent down-link in a link-down state, if said critical up-link is detected to be in a link-down state and for starting a recovery process in a host device if said active link is in the link-down state.

9. The system according to claim 8, characterised in that said monitoring device (EC) further comprises

a physical layer device (PHY) for monitoring the physical state of said up-link, and

a media access controller (MAC) for changing the state of the down-link.

10. The system according to claim 8, characterised in that the up-link of a network element (SW1, ..., SW8) is a critical up-link, if the failure of said link affects the data flow of a down-link of said network element.

11. The system according to claim 8, characterised in that the link of a network element (SW1, ..., SW8) is a dependent down-link, if there is a critical up-link between said down-link and the next network element (SW1, ..., SW8).

12. The system according to claim 8, characterised in that said monitoring device (EC) is an Ethernet controller.

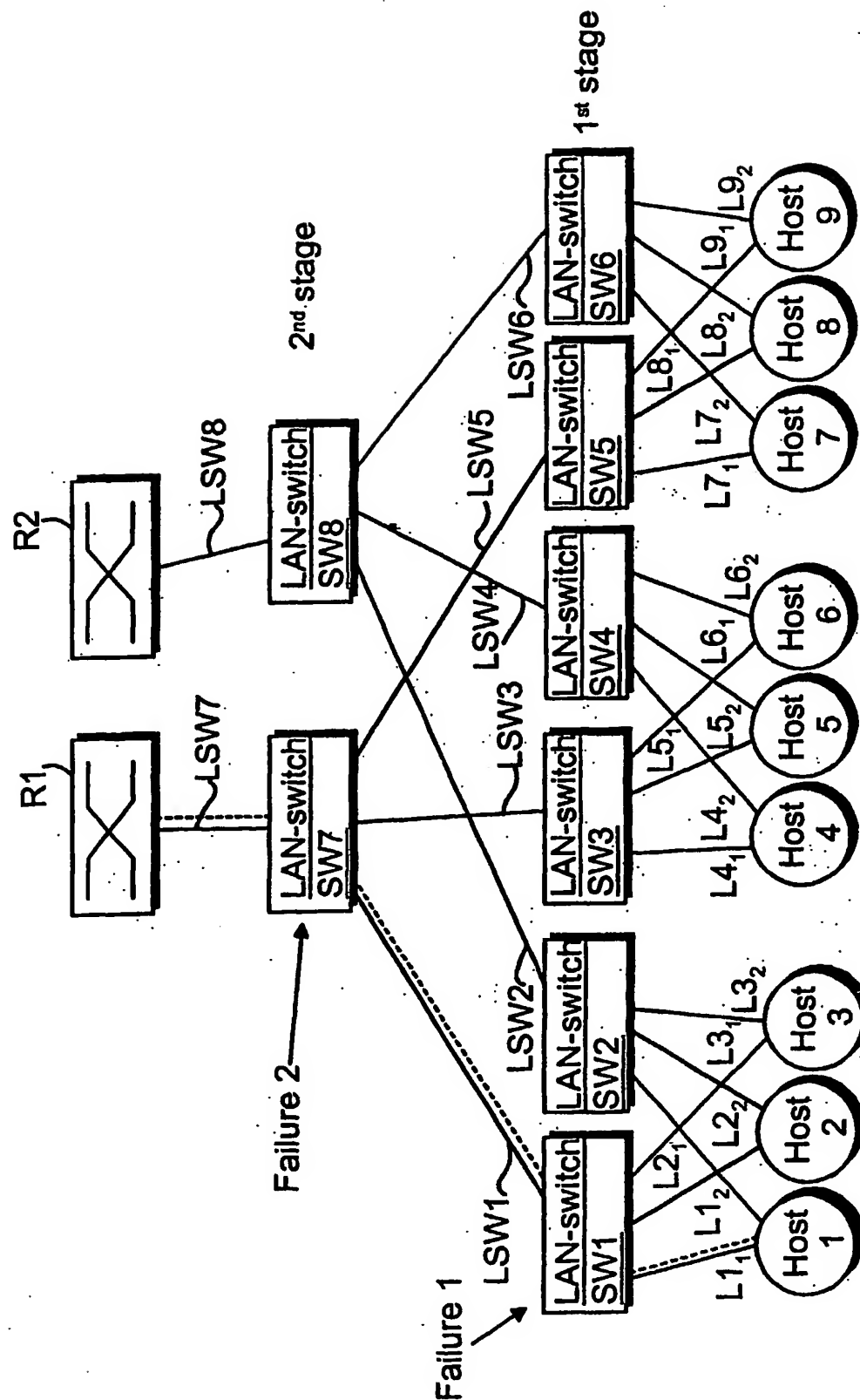
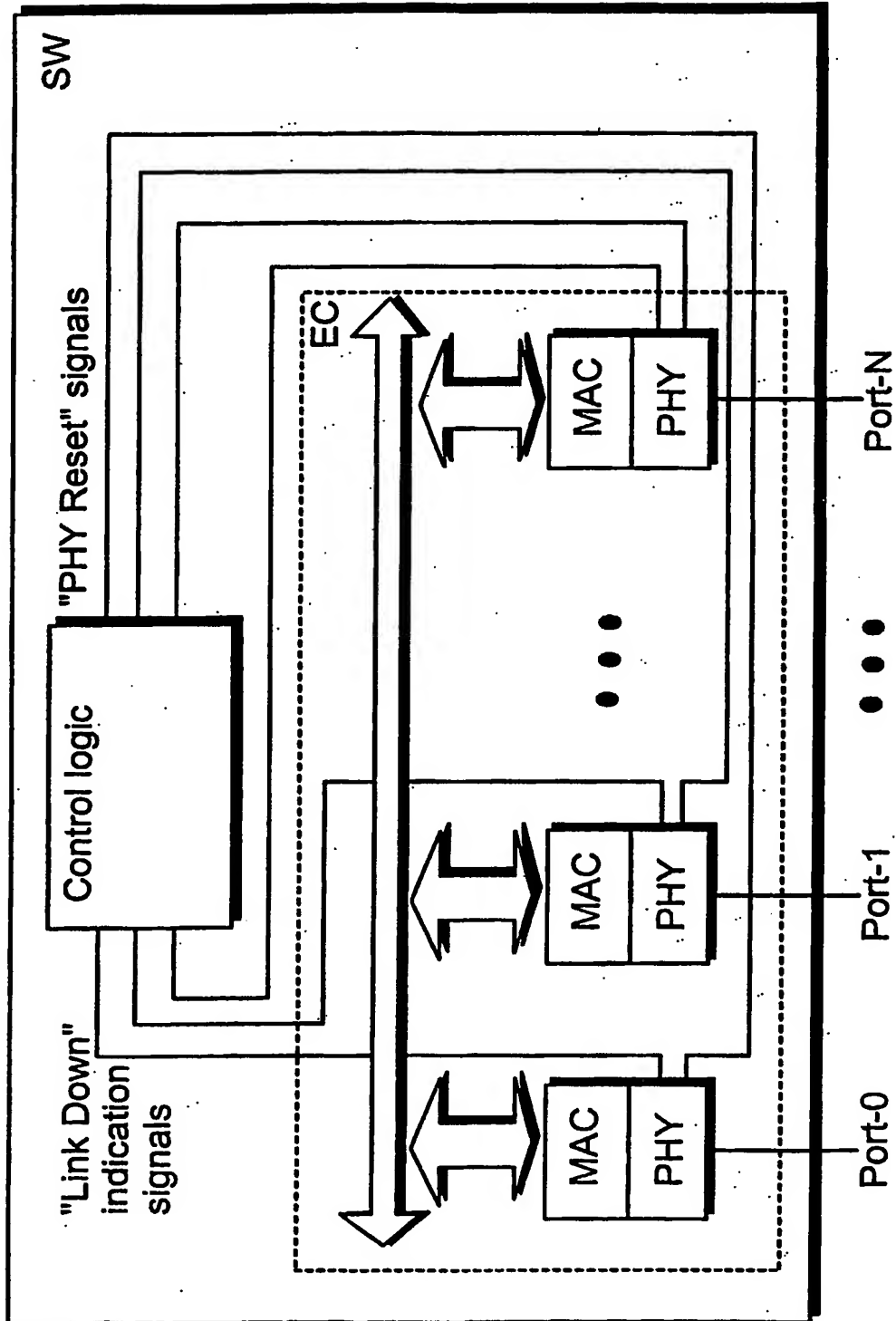


Fig. 1

Fig. 2a



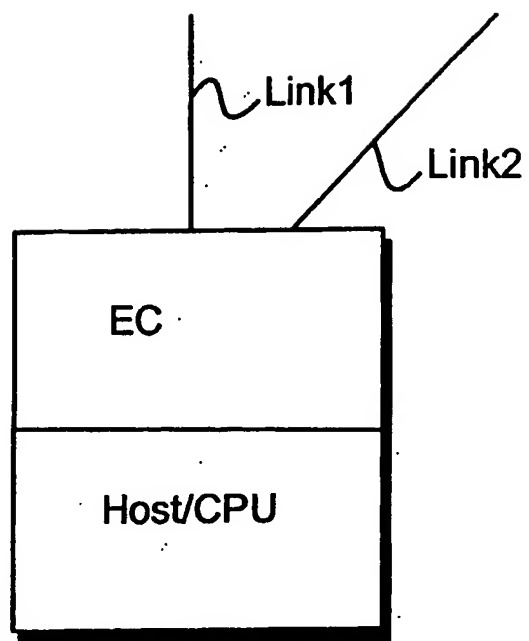


Fig. 2b

INTERNATIONAL SEARCH REPORT

International application No.

PCT/FI 02/00224

A. CLASSIFICATION OF SUBJECT MATTER

IPC7: H04L 12/24, H04L 12/56

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7: H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-INTERNAL, WPI DATA PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 0028685 A1 (3COM CORPORATION), 18 May 2000 (18.05.00), claim 18, abstract --	1-12
A	WO 9605704 A2 (BRITISH TELECOMMUNICATIONS PUBLIC LTD CO), 22 February 1996 (22.02.96), figure 6, claim 1 --	1-12
A	US 6222854 B1 (DOVE, D.J.), 24 April 2001 (24.04.01) --	1-12
A	WO 0051290 A2 (ALCATEL INTERNETWORKING, INC), 31 August 2000 (31.08.00), page 10, line 3 - line 30 --	1-12

☒ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

8 July 2002

Date of mailing of the international search report

10 -07- 2002

Name and mailing address of the ISA/

Swedish Patent Office

Box 5055, S-102 42 STOCKHOLM

Facsimile No. +46 8 666 02 86

Authorized officer

Kristoffer Ogebjær/LR

Telephone No. +46 8 782 25 00

INTERNATIONAL SEARCH REPORT

International application No.

PCT/FI 02/00224

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0957648 A2 (NORTEL NETWORKS CORP), 17 November 1999 (17.11.99), [0034] --	1-12
A	US 5732192 A (MALIN, J.T. ET AL.), 24 March 1998 (24.03.98), claim 67 -- -----	1-12

INTERNATIONAL SEARCH REPORT
Information on patent family members

10/06/02

International application No.

PCT/FI 02/00224

Patent document cited in search report			Publication date	Patent family member(s)		Publication date
WO	0028685	A1	18/05/00	AU	1717200 A	29/05/00
				US	6330229 B	11/12/01
WO	9605704	A2	22/02/96	AU	688096 B	05/03/98
				AU	3186595 A	07/03/96
				CA	2197199 A	22/02/96
				CN	1158204 A	27/08/97
				DE	69512789 D,T	27/04/00
				EP	0775427 A,B	28/05/97
				ES	2139924 T	16/02/00
				FI	970569 A	11/02/97
				HK	1013583 A	00/00/00
				JP	10504943 T	12/05/98
				NO	970627 A	11/04/97
				NZ	290910 A	26/05/97
				SG	43133 A	17/10/97
				US	5941955 A	24/08/99
US	6222854	B1	24/04/01	NONE		
WO	0051290	A2	31/08/00	AU	4004000 A	14/09/00
				CN	1341313 T	20/03/02
				EP	1171976 A	16/01/02
				AU	4003700 A	14/09/00
EP	0957648	A2	17/11/99	NONE		
US	5732192	A	24/03/98	NONE		